Kim et al. BMC Bioinformatics 2018, 19(Suppl 9):288
https://doi.org/10.1186/s12859-018-2270-7

BMC Bioinformatics

RESEARCH Open Access

Drug response prediction model using a hierarchical structural component modeling method

Sungtae Kim[1], Sungkyoung Choi[1], Jung-Hwan Yoon[2], Youngsoo Kim[3], Seungyeoun Lee[4] and Taesung Park[1,5*]

From 13th International Symposium on Bioinformatics Research and Applications (ISBRA 2017)
Honolulu, Hawaii, USA. 30 May - 2 June 2017

Abstract

Background: Component-based structural equation modeling methods are now widely used in science, business, education, and other fields. This method uses unobservable variables, i.e., "latent" variables, and structural equation model relationships between observable variables. Here, we applied this structural equation modeling method to biologically structured data. To identify candidate drug-response biomarkers, we first used proteomic peptide-level data, as measured by multiple reaction monitoring mass spectrometry (MRM-MS), for liver cancer patients. MRM-MS is a highly sensitive and selective method for proteomic targeted quantitation of peptide abundances in complex biological samples.

Results: We developed a component-based drug response prediction model, having the advantage that it first combines collapsed peptide-level data into protein-level information, facilitating subsequent biological interpretation. Our model also uses an alternating least squares algorithm, to efficiently estimate both coefficients of peptides and proteins. This approach also considers correlations between variables, without constraint, by a multiple testing problem. Using estimated peptide and protein coefficients, we selected significant protein biomarkers by permutation testing, resulting in our model for predicting liver cancer response to the tyrosine kinase inhibitor sorafenib.

Conclusions: Using data from a cohort of liver cancer patients, we then "fine-tuned" our model to successfully predict drug responses, as demonstrated by a high area under the curve (AUC) score. Such drug response prediction models may eventually find clinical translation in identifying individual patients likely to respond to specific therapies.

Keywords: Biomarkers, Component-based structural equation modeling, Drug response, Liver cancer, Multiple reaction monitoring mass spectrometry (MRM-MS), Prediction model, Sorafenib

Background

Liver cancer (hepatic cancer), is predominantly found in the tissue parenchyma, and is thus known as hepatocellular carcinoma (HCC), the most common form of liver cancer in adults. HCC can exert different growth patterns from one tumor to the next [1, 2]. In Eastern Asia, HCC is the third-most common form of cancer, and the second-leading cause of cancer death, with a worldwide total of 600,000 deaths each year [3, 4]. However, as many treatment methods have been developed for treating HCC, overall, these have shown little benefit in improving patients' prognosis [5]. More efficient treatment of HCC may lie in "personalized medicine," i.e., tailoring therapies for individual patients [6]. Such ability to classify HCC patients, with therapies optimized for specific stage and growth patterns, would reduce time and cost, and likely prolong survival.

Toward that objective, accurate prediction models are essential. Historically, methods of building cancer prediction models were based on the classification methods of linear/logistic regression, support vector machine, or

* Correspondence: tspark@stats.snu.ac.kr
[1]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 08826, South Korea
[5]Department of Statistics, Seoul National University, Seoul 08826, South Korea
Full list of author information is available at the end of the article

random forest [7–9]. While these models are effective for prediction, they do not consider any structural or hidden biological data, making it difficult to derive more meaningful biological interpretation.

Here, we built a drug response prediction model, by identifying candidate protein biomarkers, via multiple reaction monitoring-mass spectrometry (MRM-MS) technology. MRM-MS is a targeted proteomics technology that is highly selective and sensitive for quantitating targeted proteins or peptides in biological samples [10, 11]. MRM-MS can also measure several hundred protein targets per sample, simultaneously, generating consistent, precise, and reproducible datasets [12]. Consequently, MRM-MS holds high potential for biomarker discovery. Unlike other protein data, MRM-MS data is hierarchically structured.

Following mass spectrometry, our MRM-MS data consisted of 231 peptides, from 124 proteins, with each protein containing ≥1 peptide. While classical methods for prediction model building only select the best peptides as variables, to optimize prediction performance, these do not consider any biological relationship between peptides and proteins.

In this study, we built a drug response prediction model, using a component-based structural equation modeling method, based on the biological structure of MRM-MS data (e.g., peptide to protein). Structural equation modeling (SEM) is used to analyze the structural relationship between unobserved (latent) variables and observed variables. SEM can be classified as factor based SEM and component based SEM. Confirmatory factor analysis (CFA) and partial least squares path modeling (PLS-PM) analysis are the most popular methods of factor based SEM and component based SEM, respectively [13]. Our proposed model is based on generalized structured component analysis (GSCA) [14], resembling our earlier derivation of a pathway-based approach. That analysis, using hierarchical components of collapsed rare variants (PHARAOH), uses a hierarchical structure of pathways and genes [15].

Using latent variables, we can collapse multiple peptides into a structured form of proteins that they comprise of, providing more feasible biological explanations of the results. In addition to hierarchical structure, we further showed that HisCoM effectively cover protein-level analysis, taking all peptides into account simultaneously. Moreover, for real biological data analysis, using MRM-MS, we discovered possible protein biomarkers associated with patients' response to the multiple tyrosine kinase inhibitor sorafenib (Nexavar®) [16]. Sorafenib is known as effective and safe drug for recovering liver cancer (hepatocellular carcinoma) patients not only Asian-Pacific region but also in the North American region [17, 18]. Using these protein biomarkers, we then evaluated the performance of our drug response prediction model. Additionally, we compared the performance of our prediction model, using area under the curve (AUC) scores, to performances by generalized linear models of logistic regression, without ridge parameters, and logistic regression, with ridge parameters. Furthermore, through extensive simulation studies, we compared the performance of our proposed method with other logistic regression methods. For hierarchical structuring, in this case, for proteins with multiple peptides, our HisCoM was shown to perform better than logistic regression, as assessed by AUC scores.

Methods

Preparing samples and materials

Hepatocellular carcinoma (HCC) patient serum samples ($n = 115$) were collected at Seoul National University Hospital, from 2013 to 2015 [19]. Upon diagnosis of liver cancer, patients were placed on a regimen using the tyrosine kinase inhibitor Sorafenib (Nexavar®, Bayer, Inc., Whippany, NJ, USA). Patients' tumor sizes were first examined immediately following HCC diagnosis, at the start of hospital admission. Six weeks after first diagnosis (sufficient time to see a response), patients' tumors were again measured, by contrast-enhanced computed tomography or magnetic resonance imaging, and staged according to the standardized Modified Response Evaluation Criteria in Solid Tumors (mRECIST) [20]. After the second examination, patients were divided into two groups, based on positive and negative drug responses. The positive drug response group consisted of patients with complete response (CR), partial response (PR), or stable disease (SD), according to mRECIST [20]. CR and PR responses were diagnosed when the tumor size was reduced after 6 weeks. Also, SD was diagnosed when the size of the tumor remained stable, from the first to second visit. On the other hand, the negative drug response group consisted of patients with progressive disease (PD), wherein the size of their tumors increased, from first diagnosis to 6 weeks later. The study protocol was approved by the Institutional Review Board of Seoul National University Hospital (IRB No. 0506–150-005), and written, informed consent was obtained from each patient or legally authorized representative.

Among all 115 patients (101 men and 14 women), 40 patients (37 men and 3 women) were grouped into the positive drug response group, and 75 (64 men and 11 women) were grouped into the negative drug response group. From each patient's serum, data for 231 peptides was generated by multiple reaction monitoring mass spectrometry (MRM-MS), a highly sensitive and selective method for targeted quantitation of peptide abundances, in complex biological samples [21]. Here, the 231 peptides can represent

124 proteins. Since the MRM-MS technique measures the quantity of targeted peptides in patients' serum, we used the log2-transformed ratio of light peptide intensity to heavy peptide intensity. Light peptide intensity represented the quantity of peptides from specific patient's blood, as measured by MRM-MS, while heavy peptide intensity represented the quantity of artificially built, same sequences as the light peptides, but using heavier isotope elements, also as measured by MRM-MS. The software Skyline was used to measure the intensity of light and heavy peptides by MRM-MS [22]. Demographic information, such as age and sex, were also available. The range of age varied from 34 to 84, with 101 male and 14 female samples.

Constructing the drug response model

The overall schematic procedure is shown in Fig. 1. At the beginning, we selected protein level biomarkers by HisCoM method with 1000 permutation test, for possible prediction of sorafenib response, using MRM-MS data. Second, we constructed prediction models, via a component-based structural equation-modeling method. Finally, we evaluated the constructed drug response prediction models' performances, by AUC scores.

An example of our proposed drug response prediction model is shown in Fig. 2. This model combines collapsed peptide-level MRM-MS data into protein-level information, and efficiently estimates both peptide and protein coefficients. In this example, two proteins were involved ($K = 2$), and each protein consisted of two or three peptides ($T_k = 2$ and 3). Weight (w) and path coefficients (β) were estimated using alternating least squares [23].

Here, suppose that there are K proteins, and the k^{th} protein contains T_k peptides, for $k = 1,\ldots,K$. To estimate parameters, the following penalized log likelihood function was maximized. The y_j represent the drug response group, based on mRECIST: $y_j = 0$ for a negative response, and $y_j = 1$ for a positive response. Let y_j is distributed independently with a mean of $\mu_j = \mathrm{E}[y_j]$ and η_j is defined as $\eta_j = g(\mu_j)$ by a logit link function g. Then, we can derive a penalized log likelihood function with dispersion parameter δ and canonical parameter γ_i as following:

$$\varphi_1 = \sum_{j=1}^{N} \log P\left(y_j; \gamma_i, \delta\right) - \frac{1}{2}\lambda_{pep}\sum_{k=1}^{K}\sum_{t=1}^{T_k} w_{kt}^2 - \frac{1}{2}\lambda_{prot}\sum_{k=0}^{K}\beta_k^2 \tag{1}$$

Here, λ_{prot} and λ_{pep} are the ridge parameters for proteins and peptides represent as "tuning" parameters, respectively: one for the peptides within a protein and the other for the proteins themselves.

Let $\boldsymbol{w}_k = [w_{k1}, \cdots, w_{kT_k}]$, $\boldsymbol{\beta} = [\beta_0, \beta_1, \cdots, \beta_K]$, $\boldsymbol{F} = [f_1, \cdots, f_N]$, and $\boldsymbol{f}_j = [1, f_{j1}, \cdots, f_{jK}]$ where $f_{jk} = \sum_{t=1}^{T_k} x_{jki} w_{ki}$. We define x_{jki}as the quantity of i^{th} peptide of the k^{th} protein in sample j. The w_{ki} as a weight coefficient of i^{th} peptide of the k^{th} protein. Also, the β_k as a path coefficient of k^{th} protein. Maximizing the eq. (1), via iteratively reweighted least squares, is identical to minimizing the following penalized least squares eq. (2):

$$\begin{aligned}\varphi_2 &= \sum_{j=1}^{N} v_j\left(z_j - \sum_{k=0}^{K} f_{jk}\beta_k\right)^2 + \lambda_{pep}\sum_{k=1}^{K}\sum_{t=1}^{T_k} w_{kt}^2 + \lambda_{prot}\sum_{k=0}^{K}\beta_k^2 \\ &= (\boldsymbol{z}-\boldsymbol{F}\boldsymbol{\beta})'\boldsymbol{V}(\boldsymbol{z}-\boldsymbol{F}\boldsymbol{\beta}) + \lambda_{pep}\sum_{k=1}^{K}(\boldsymbol{w}_k'\boldsymbol{w}_k) + \lambda_{prot}(\boldsymbol{\beta}'\boldsymbol{\beta})\end{aligned} \tag{2}$$

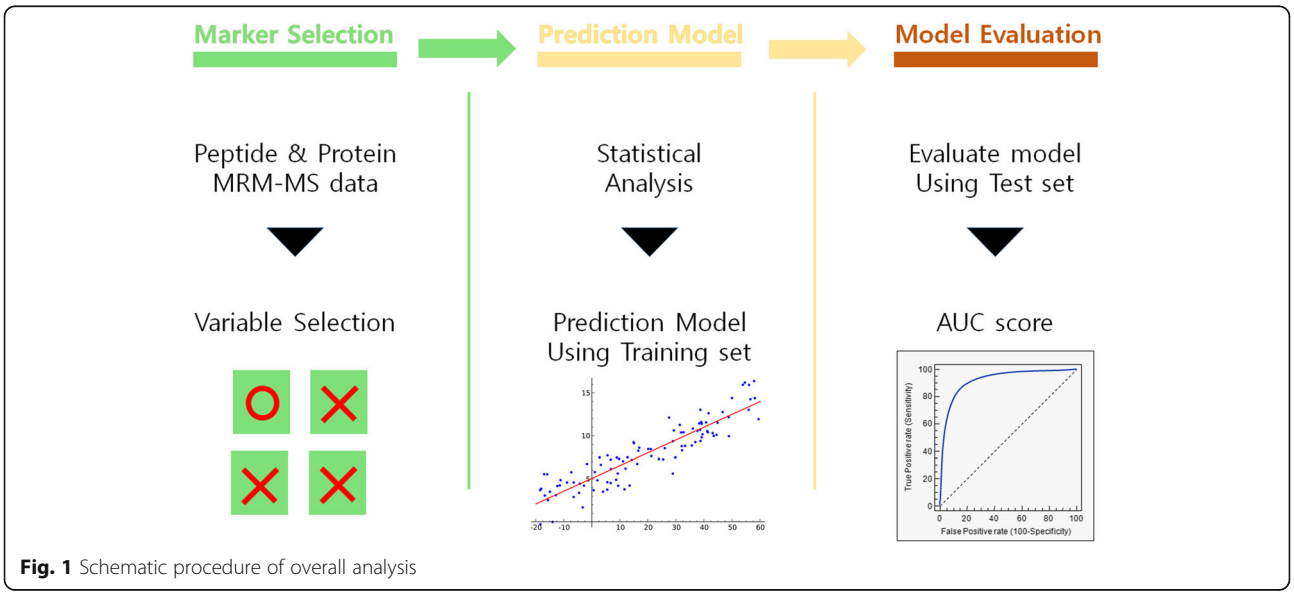

Fig. 1 Schematic procedure of overall analysis

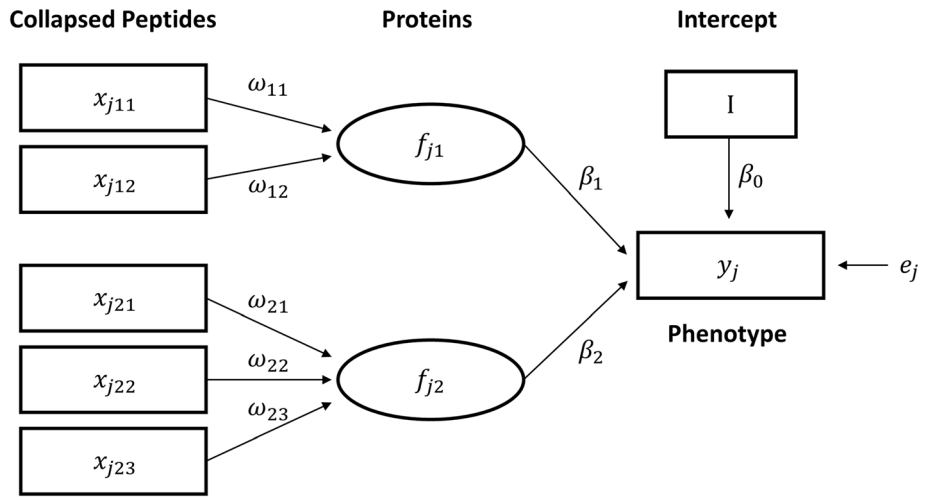

Fig. 2 Example of proposed component-based structured equation model

with respect to $\boldsymbol{w}_k$ and $\boldsymbol{\beta}$ [15, 24]. Here, $\boldsymbol{V}$ is an N by N diagonal matrix with elements $v_j = \frac{(\partial_{\mu_j}/\partial_{\eta_j})^2}{\tau_j}$. τ_j is the variance function evaluated at μ_j. The $\boldsymbol{z}$ is the adjusted response variable and an $N \times 1$ vector with elements $z_j = \eta_j + \frac{(y_j - \mu_j)}{v_j}$ [25].

After estimating the w_{ki} and β_k coefficients, we constructed a drug response prediction model for $\pi_j = P[y_j = 1] = \mu_j$, as follows, after standardization of x_{jki} The coefficients of age and sex are also estimated by maximizing the log-likelihood function simultaneously while penalizing the coefficients of peptides. In our final prediction model, the beta coefficients of age and sex are fixed across the individuals.

$$log\left(\frac{\pi_j}{1-\pi_j}\right) = \beta_0 + \sum_k \left(\sum_i x_{jki} w_{ki}\right)\beta_k + AGE_j\beta_{age} + SEX_j\beta_{sex} \quad (3)$$

$$= \beta_0 + \sum_k f_{jk}\beta_k + AGE_j\beta_{age} + SEX_j\beta_{sex} \quad (4)$$

j: individual samples ($j = 1, \cdots, 115$)
k: proteins ($k = 1, \cdots, 124$)
i: peptides ($i = 1, \cdots, 231$)

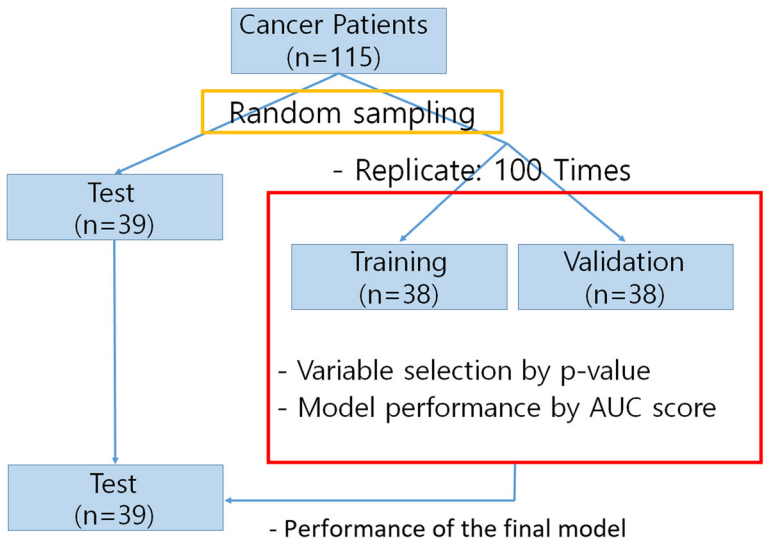

Fig. 3 Sample separation for training, validation, and test sets

When our final drug response prediction model was constructed, we evaluated its performance by area under the receiver operating characteristic curve (AUC) score, based on the training, validation, and test sets. The approach for separating the training set from the validation and test sets, is depicted in Fig. 3. First, we then randomly selected 39 out of 115 samples (35 men and 4 women) as a test set, excluded from the modelling process, while assessing the remaining 76 samples. This test set will be used to measure the performance of our final drug response prediction model. The ratio of positive responses to negative responses was sustained (14 positive responses and 25 negative responses). The range of age in test set was distributed from 41 to 84. The remaining 76 samples were randomly (without replacement) divided into training and validation set. For a fair comparison, the ratio of positive responses to negative responses was retained (13 positive responses and 25 negative responses). The concept of a sample separation process was based on a previously developed intraductal papillary mucinous neoplasm (IPMN) patient prediction model [26].

Lastly, our drug response prediction model was compared to the generalized linear model with a binary response (GLM), and the generalized linear regression with a binary response via ridge parameter (GLMwR) methods. All the analyzes were calculated and computed via software R (Version R 3.2.3) [27].

Simulation design

For the simulation study, we designed two models: the first model composed of two significant proteins and the second model with both significant and nonsignificant protein in the presence of a hierarchical structure of MRM-MS data (e.g., peptide to protein). Let the first simulation model contain JCHAIN and RBP4 with parameters estimated by HisCoM. Note that JCHAIN was a significant protein (p-value: 0.0142), with 3 peptides, and RBP4 was also a significant protein (p-value: 0.0031), with 2 peptides. The simulation model is given by

Table 1 *P*-values for our 6 candidate biomarkers: APOC4, CD163, CD5L, JCHAIN, SERPING1, and RBP4, based on MRM-MS data

Protein	*P*-value
APOC4	0.0061
CD163	0.0112
CD5L	0.0031
SERPING1	0.0102
JCHAIN	0.0142
RBP4	0.0031

$$\log\left(\frac{\pi_j}{1-\pi_j}\right) = \beta_0 + \left(\sum_{i=1}^{3} x_{ji}\omega_i\right)\beta_{IGJ} + \left(\sum_{i=4}^{5} x_{ji}\omega_i\right)\beta_{RET4} + AGE_j\beta_{age} + SEX_j\beta_{sex} \quad (5)$$

For the Simulation model 2, we assume the true model contains RBP4 and APOA1, with parameters estimated by HisCoM. Note that RBP4 was a significant protein (p-value: 0.0031), with 2 peptides, and APOA1 was a nonsignificant protein (p-value: 0.4794), with 7 peptides. The second simulation model is given by

$$\log\left(\frac{\pi_j}{1-\pi_j}\right) = \beta_0 + \left(\sum_{i=1}^{2} x_{ji}\omega_i\right)\beta_{RET4} + \left(\sum_{i=3}^{9} x_{ji}\omega_i\right)\beta_{APOA1} + AGE_j\beta_{age} + SEX_j\beta_{sex} \quad (6)$$

In this case, $x_{j,\ 1}$ represents the j^{th} individual's peptide data ($x_{j,1}$, $x_{j,2}$, $\cdots$, $x_{j,9}$). From the estimated βs and ωs,

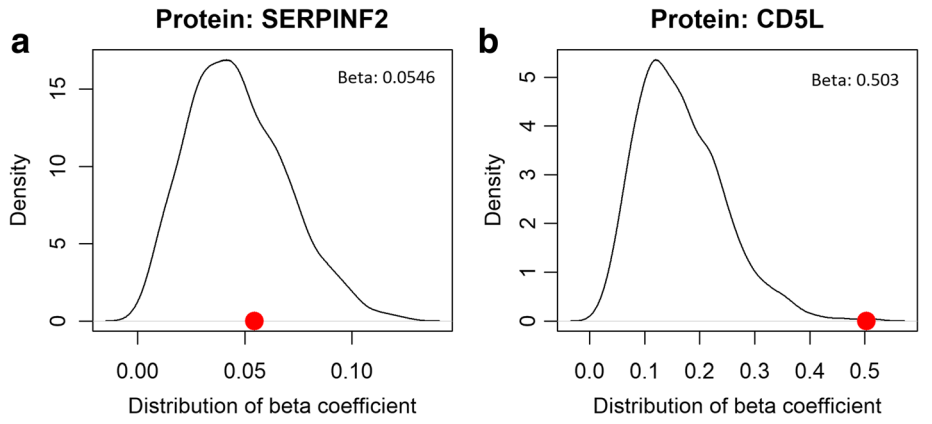

Fig. 4 Estimation of beta coefficients (path coefficients) for each protein example. **a** non-significant proteins; **b** significant proteins. Red dot indicates the estimated path coefficient of the protein

Table 2 AUC score comparison between HisCoM, GLM, and GLMwR drug response models using single candidate protein

Protein	HisCoM	GLM	GLMwR
APOC4	0.617	0.611	0.611
CD163	0.697	0.703	0.697
CD5L	0.860	0.883	0.897
SERPING1	0.837	0.857	0.846
JCHAIN	0.717	0.709	0.700
RBP4	0.803	0.829	0.826

Table 3 AUC score comparison between HisCoM, GLM, and GLMwR drug response models using double candidate proteins

Protein	HisCoM	GLM	GLMwR
APOC4_CD163	0.851	0.837	0.837
APOC4_CD5L	0.886	0.851	0.880
APOC4_SERPING1	0.834	0.871	0.846
APOC4_JCHAIN	0.886	0.814	0.897
APOC4_RBP4	0.786	0.794	0.789
CD163_CD5L	0.866	0.883	0.897
CD163_SERPING1	0.894	0.917	0.914
CD163_JCHAIN	0.731	0.731	0.729
CD163R_RBP4	0.900	0.880	0.889
CD5L_SERPING1	0.923	0.923	0.934
CD5L_JCHAIN	0.854	0.931	0.886
CD5L_RBP4	0.917	0.937	0.926
SERPING1_JCHAIN	0.940	0.943	0.946
SERPING1_RBP4	0.871	0.891	0.897
JCHAIN_RBP4	0.929	0.911	0.931

derived from the data, we estimated π_1, π_2,$\cdots$, π_{115}. Then, the responses were generated from the Bernoulli distribution $B(1, \pi_j)$, for $j = 1, 2, \cdots, 115$. We then constructed HisCoM, GLM, and GLMwR drug response prediction models, using MRM-MS peptide data ($x_{j,1}$, $x_{j,2}$,$\cdots$, $x_{j,\ 9}$), to generate response variables. For each simulation model, we measured the AUC score. Using the same estimated values of π_1, π_2,$\cdots$, π_{115}, we repeated the whole process 1000 times, and obtained 1000 AUC scores for each of the HisCoM, GLM, and GLMwR models. We then calculated the mean of the 1000 AUC scores, based on those models.

Results

Biomarker discovery for the drug response prediction model

To evaluate our model, at the beginning, we randomly selected 39 out of 115 samples, as a separate, test set, to evaluate the overall performance of the final drug response prediction model. We performed cross-validation analysis using remaining 76 samples. The dataset was randomly divided into training/validation sets (38 samples for each set). From the training data set, the significant proteins were selected based on p-values. Then the prediction model was build, and its AUC score was computed from the validation set. We repeated this cross-validation 100 times. Through 100 cross-validation, we evaluated whether the significant proteins were selected repeatedly by HisCoM. Also, using the estimated path coefficients by training set, we evaluated the performance of the prediction model from the validation set.

The significances of the protein path coefficients were then determined, using a 1000 permutation test, for each

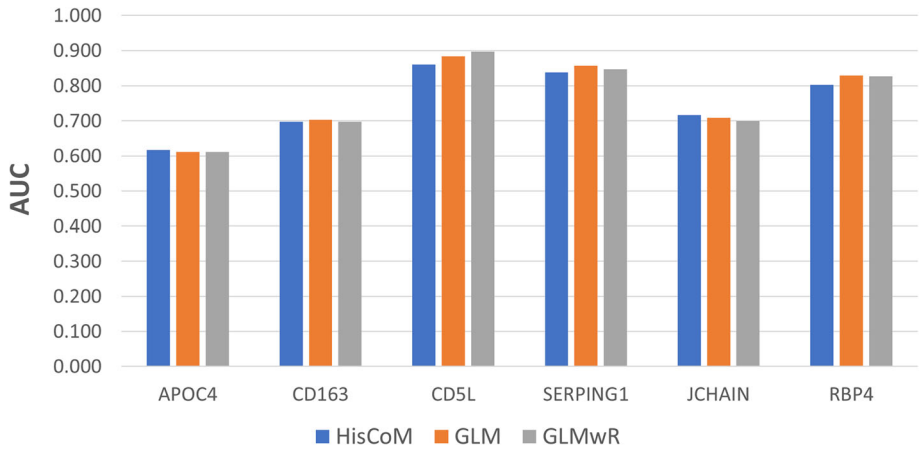

Fig. 5 AUC score comparison between HisCoM, GLM, and GLMwR drug response prediction models using single candidate protein

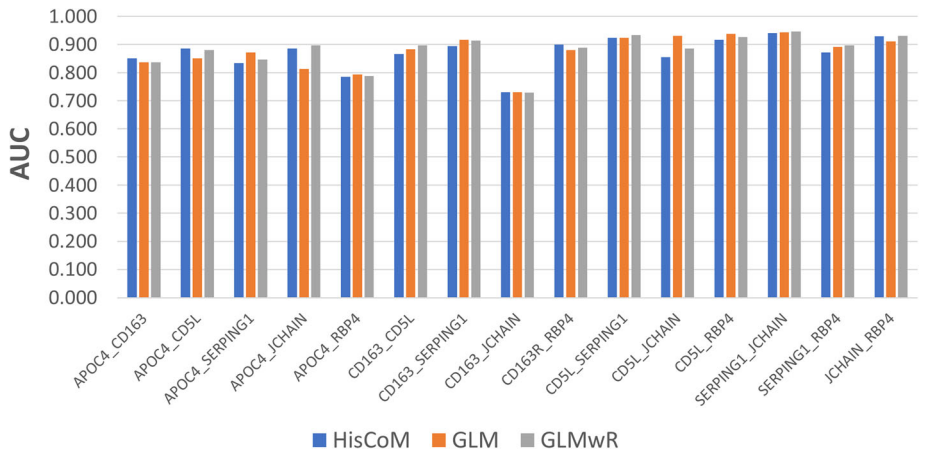

Fig. 6 AUC score comparison between HisCoM, GLM, and GLMwR drug response prediction models using double candidate proteins

replicate. The permutation test was performed by shuffling drug response variables randomly across subjects while retaining the ratio of positive response to negative response and then estimating the path coefficients. Using path coefficients estimated by 1000 permutation test, we construct distribution of each protein's path coefficient. Through comparing the path coefficient value of the original data with those from the permuted data, p-values were computed for each protein, with the significant ($P < 0.05$) ones being selected for further analysis.

Figure 4 shows the null distribution of path coefficients (β_k), derived from 1000 permutation tests. Figure 4a show the case of non-significant protein, while the Fig. 4b does the case of significant protein. The red dots indicate the estimated path coefficients from the data. In our analysis, the path coefficients (β_k) of six proteins were significant.

During the processing of the training/validation sets, with 76 samples, we repeated this process 100 times, to check the consistency of possibly significant proteins. Since our method uses two ridge parameters, we defined the same tuning parameter values as 10 for peptides and proteins, for computational efficiency. As a result, we selected the top 6 significant proteins (APOC4, CD163, CD5L, JCHAIN, SERPING1, and RBP4), which were repeatedly selected as significant by the process of 100 replications. We noted that these six proteins were previously identified as possible proteomic biomarkers, for hepatocellular carcinoma [28–30].

We then repeated the process once more, with only those 6 proteins, as MRM-MS data, for more accurate estimation of p-values and path coefficients, for the drug response prediction model. We next calculated p-values and path coefficients. In Table 1, p-values are shown for the six selected proteins.

Using the selected 6 proteins, we constructed a drug response prediction model, with estimated w and β values. We also constructed different prediction models, limiting the number of proteins. In this case, we constructed models using 1 of the 6 proteins, 2 of the 6, 3 of the 6, and all six. For all these models, age and sex were considered as covariates (see eqs. 3 and 4, below). All the analyzes were calculated and computed via software R (Version R 3.2.3) [27].

Table 4 AUC score comparison between HisCoM, GLM, and GLMwR drug response models using triple candidate proteins

Protein	HisCoM	GLM	GLMwR
APOC4_CD163_CD5L	0.920	0.883	0.920
APOC4_CD163_SERPING1	0.880	0.886	0.877
APOC4_CD163_JCHAIN	0.917	0.894	0.914
APOC4_CD163_RBP4	0.886	0.866	0.874
APOC4_CD5L_SERPING1	0.920	0.937	0.931
APOC4_CD5L_JCHAIN	0.920	0.891	0.926
APOC4_CD5L_RBP4	0.897	0.886	0.906
APOC4_SERPING1_JCHAIN	0.946	0.960	0.943
APOC4_SERPING1_RBP4	0.877	0.863	0.869
APOC4_JCHAIN_RBP4	0.920	0.866	0.929
CD163_CD5L_SERPING1	0.940	0.923	0.949
CD163_CD5L_JCHAIN	0.869	0.929	0.914
CD163_CD5L_RBP4	0.951	0.943	0.949
CD163_SERPING1_JCHAIN	0.957	0.957	0.954
CD163_SERPING1_RBP4	0.923	0.929	0.911
CD163_JCHAIN_RBP4	0.954	0.909	0.954
CD5L_SERPING1_JCHAIN	0.923	0.937	0.940
CD5L_SERPING1_RBP4	0.946	0.957	0.949
CD5L_JCHAIN_RBP4	0.943	0.937	0.940
SERPING1_JCHAIN_RBP4	0.946	0.943	0.943

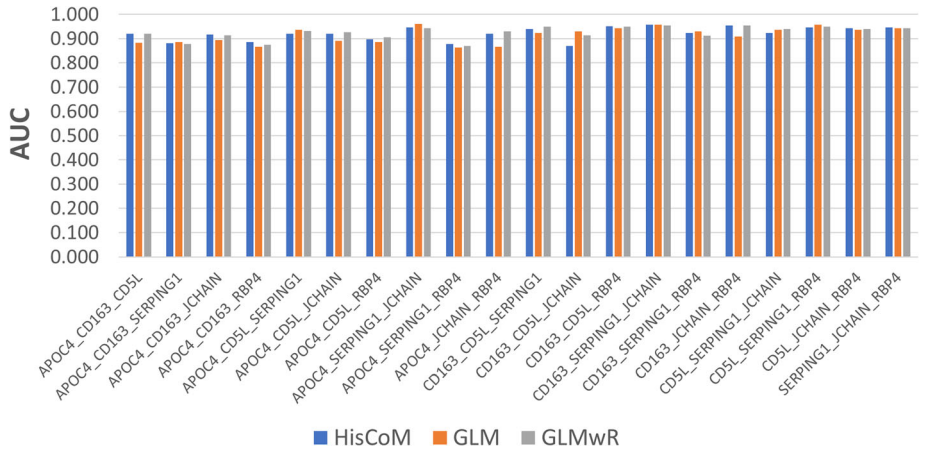

Fig. 7 AUC score comparison between HisCoM, GLM, and GLMwR drug response prediction models using triple candidate proteins

Model evaluation by AUC results

With the selected proteins, we first constructed a sorafenib drug response prediction model given in eq. (4), using HisCoM. The performance of the drug response prediction models was measured by AUC scores. In this case, the numbers k and i varied, depending on the number of proteins in the model.

Table 2 shows the AUC score of our single protein prediction model, compared to a corresponding the generalized linear model with a binary response (GLM), and the generalized linear model with a binary response via ridge parameters (GLMwR). The performance of the single protein prediction models showed similar AUC scores, across all three different statistical methods, while the AUC scores, for each individual protein, varied from 0.60 to 0.90. Figure 5 provides a visual comparison of AUC scores, by different single protein statistical models.

The prediction model, using 2 of the 6 proteins, had higher AUC scores, compared to the single protein models, across all three different statistical methods. Table 3 shows the AUC scores for the models with 2 of the 6 proteins. The AUC scores across each statistical model varied from 0.73 to 0.95, higher than those for the single protein prediction models. Figure 6 shows a visual comparison, of AUC scores, by different two-protein statistical models. The HisCoM's AUC score was similar to those of GLM and GLMwR, but had higher performance or lower performance, depending on the combination of proteins. The best performing protein combination, across each HisCoM, GLM, and GLMwR statistical model, was the combination of SERPING1 and JCHAIN.

Similarly, most of the prediction models, with 3 of the 6 proteins, scored higher than 0.9 AUC, using all three modeling methods. Table 4 shows the AUC scores for each

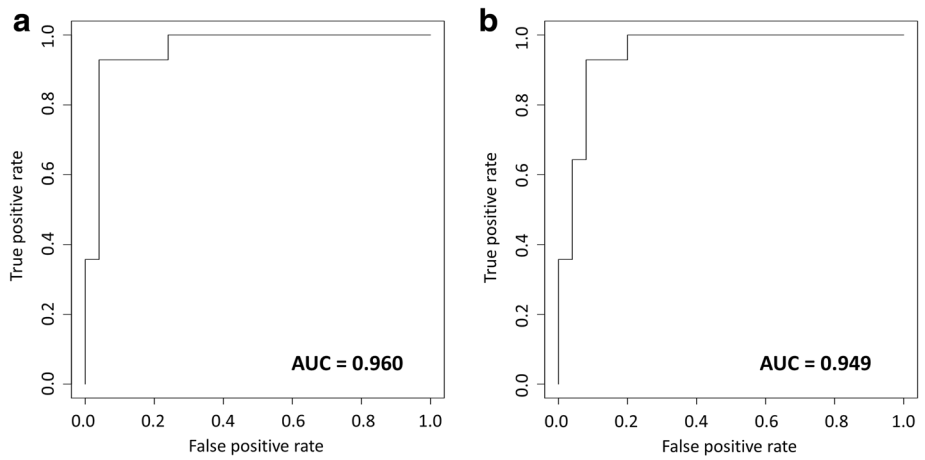

Fig. 8 AUC score, using all 6 proteins, for each model. **a** HisCoM. **b** Generalized linear model with ridge parameter

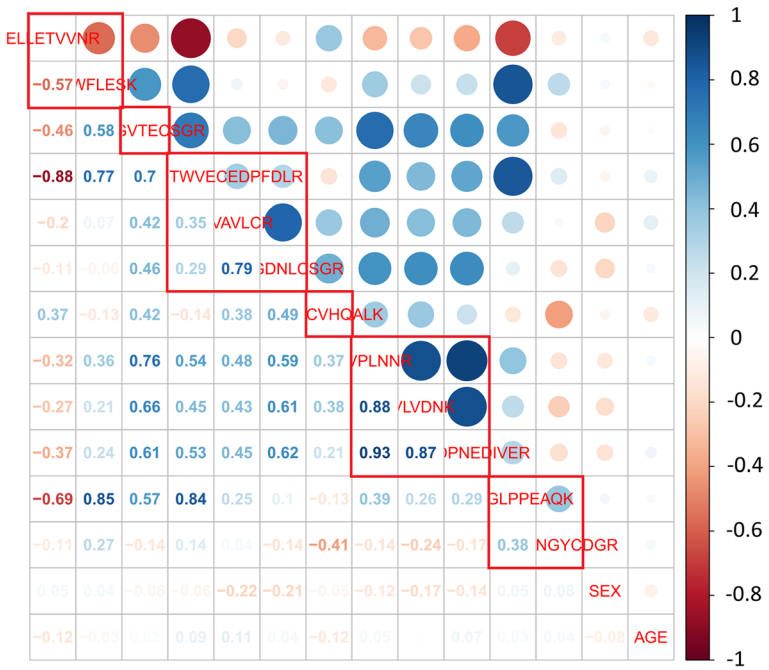

Fig. 9 Correlations between peptides. Each red square box represents the peptides within the same protein

model using exhaustive combinations of the three proteins, varying from 0.86 to 0.95. Figure 7 shows a visual comparison of AUC scores, by different 3-protein statistical models.

Using all 6 proteins, we also constructed HisCoM drug response prediction models, with estimated ω and β as the covariates age and sex, respectively. In Fig. 8, the AUC score for our HisCoM model was 0.96, using the validation set. At first, we tried to compare our prediction model to the generalized linear model with a binary response. However, the latter had a convergence problem, due to high correlation among peptides, as shown in Fig. 9. To resolve this problem, we fit the logistic regression model with a ridge penalty, using the "GLMNET" R Package. The result is shown in Fig. 8, and the AUC score for the generalized linear model with a binary response via ridge parameter (GLMwR), was 0.949, using the same validation set. As a result, our HisCoM had a slightly better AUC score (0.96), compared to that of GLMwR (0.949).

Simulation results

The performance of Simulation model 1 (JCHAIN + RBP4) results, the mean AUC scores of HisCoM, GLM and GLMwR are shown in Table 5. The mean AUC score of HisCoM was 0.8362. Figure 10 shows the range of 1000 AUC scores, as depicted by box plots, with respect to each statistical method. It shows that the HisCoM performed better than others. The performance of Simulation model 2 (RBP4 + APOA1) results, the mean AUC scores of HisCoM, GLM and GLMwR, are shown in Table 6. These results show that HisCoM had the highest mean AUC score, compared to the two other statistical methods. The mean AUC score of model 2 by HisCoM was 0.7270, while the means of the other statistical methods were less than 0.7. Also, Fig. 11 shows the range of 1000 AUC scores, as depicted by box plots, with respect to each statistical method model.

In summary, both simulation model results show that HisCoM was the best performing model, compared to the other statistical methods, when there exists a hierarchical structure of MRM-MS data (e.g., peptide to protein).

Discussion

In this study, we developed a prediction model for tumor response to the multiple tyrosine kinase inhibitor sorafenib (Nexavar®), for liver cancer patients [16], using a component-based structural equation modeling method. We used HisCoM to construct the model, for Korean hepatocellular carcinoma (HCC) patients, using MRM-MS proteomic data, including some demographic variables. HisCoM fit the whole data set at once. In this case, we measured 231 peptides' weights, and 124

Table 5 Mean AUC scores of HisCoM-based Simulation model 1

Methods	Mean AUC
HisCoM	0.8362
GLM	0.8142
GLMwR	0.8018

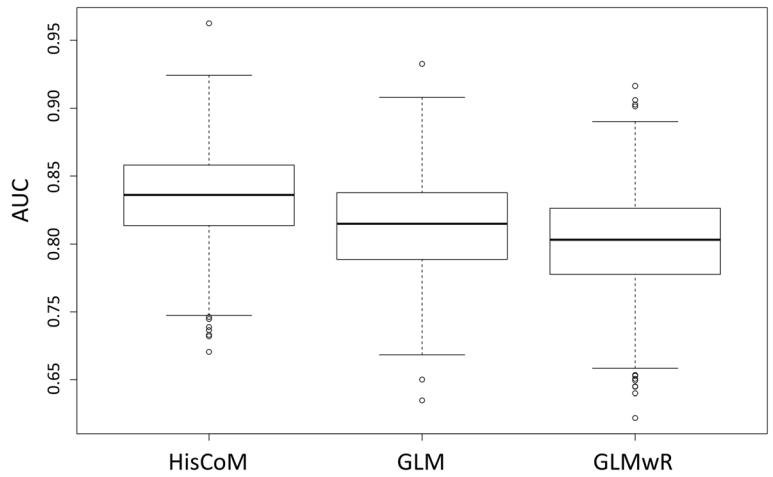

Fig. 10 Box plots of ranges of 1000 AUC scores of HisCoM-based simulation data, compared to other models: Simulation model 1

protein's path coefficients, to the drug response variable, all at once. The positive or negative drug response variables were defined by tumor responses according to mRECIST [20]. Thus, this model can be used for large-scale, structured data, with marker selection (as well as model building), simultaneously. The second, and most important advantage of HisCoM, is that it generates latent variables, which are not directly observed, while collapsing other (observed) variables. For example, our HisCoM combines several collapsed peptides' MRM-MS data, into several proteins, as latent variables. Unlike other classical methods, such as linear/logistic regression, support vector machine, and random forest, our HisCoM approach considers peptide-to-protein computational structure, and peptide-to-protein biological structure. In the analysis, we found 6 possible protein biomarkers that significantly associate with sorafenib drug response. On the other hand, other classical prediction modeling methods do not consider structure of biological information. Using peptide-level data, we found significant proteins, as possible biomarkers, for building a sorafenib response prediction model for liver cancer patients. The overall work flow, with our statistical analysis, using a HisCoM schema, can be accurately applied not only to other cancers, but also to most any large-scale structured data.

Conclusions

From possible biomarker selection, to AUC performance test scores, through a model-building process, we compared the performance of our model, constructed using a HisCoM method, to other classical statistical methods such as generalized linear models, using logistic regression (alone) or logistic regression with ridge parameters. For possible drug response biomarkers, 6 significant proteins were statistically selected, using *p*-values, as computed by permutation tests: APOC4 (*p*-value: 0.0061), CD163 (*p*-value: 0.0112), CD5L (*p*-value: 0.0031), JCHAIN (*p*-value: 0.0102), SERPING1 (*p*-value: 0.0142), and RBP4 (*p*-value: 0.0031). All six of these proteins were previously reported as possible biomarkers for hepatocellular cancer (HCC) [31–33]. Of these, CD5L is the best-known HCC biomarker [28]. For the single protein model, using HisCoM, the AUC scores varied from 0.60 to 0.90, depending on the specific protein. For modeling combinations of 2 of the 6 proteins, by HisCoM, the AUC scores varied from 0.73 to 0.95, showing increased performance, compared to single-protein prediction models. On the other hand, AUC scores varied from 0.86 to 0.95, for the 3-protein model, by HisCoM. Finally, using all six of the above-mentioned proteins in the model, we successfully constructed a drug response prediction model using 1-, 2-, 3-, or all six-protein models. By comparing our model's performance with the generalized linear model with a binary response via ridge penalization, the performance of our six-protein HisCoM prediction model was an AUC score of 0.96, slightly better than the generalized linear model with a binary response via ridge parameter, for the 6-protein panel, with an AUC score of 0.949 (Fig. 8). Thus, both the HisCoM and GLMwR methods had high AUC scores. Overall, we conclude that our model was marginally superior to the classical model types.

For future research, we can apply this overall prediction model-building approach, using HisCoM, to other

Table 6 Mean AUC scores of HisCoM-based Simulation model 2

Methods	Mean AUC
HisCoM	0.7270
GLM	0.6515
GLMwR	0.6812

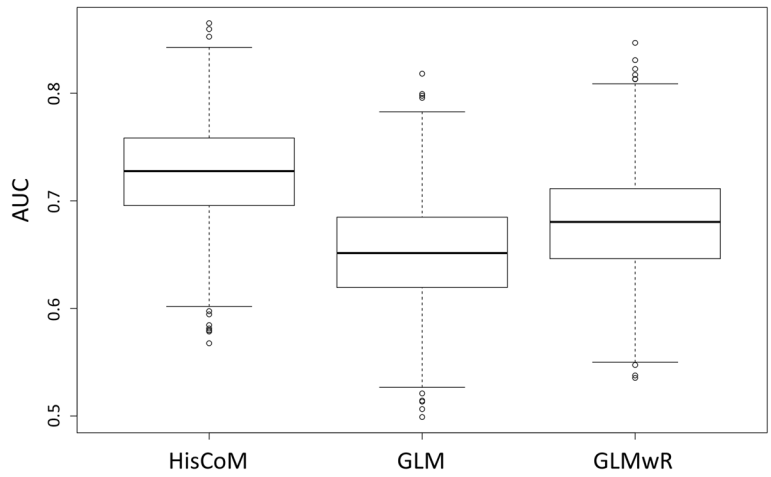

Fig. 11 Box plots of ranges of 1000 AUC scores of HisCoM-based simulation data, compared to other models: Simulation model 2

cancer data especially derived from MRM-MS platform. Since these potential biomarkers were identified in patients' serum, these could be obtained by a minimally invasive procedure (e.g., as compared to biopsies, lumbar puncture, etc.). Such models could ultimately assist physicians in discerning which therapies might be effective, for individual patients.

Abbreviations

AUC: Area under the curve; CR: Complete response; GLM: Generalized linear model; GLMwR: Generalized linear model with ridge parameters; GSCA: Generalized structured component analysis; HCC: Hepato cellular carcinoma; HisCoM: Hierarchical structural Component Models; mRECIST: Modified Response Evaluation Criteria in Solid Tumors; MRM-MS: Multiple reaction monitoring mass spectrometry; PD: Progress disease; PHARAOH: Pathway-based approach using HierArchical components of collapsed RAre variants Of High-throughput sequencing data; PR: Partial response; SD: Stable disease

Acknowledgements

The abridged abstract of this work was previously published in the Proceedings of the 13th International Symposium on Bioinformatics Research and Applications (ISBRA 2017), Lecture Notes in Computer Science: Bioinformatics Research and Applications [34].
The authors thank Dr. Curt Balch, for extensive English editing.

Funding

This work was supported by the Bio-Synergy Research Project (2013M3A9C4078158) of the Ministry of Science, ICT and Future Planning through the National Research Foundation and by grants from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI15C2165, HI16C2037).
Publication of this article was funded by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI16C2037).

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

About this supplement

This article has been published as part of *BMC Bioinformatics* Volume 19 Supplement 9, 2018: Selected articles from the 13th International Symposium on Bioinformatics Research and Applications (ISBRA 2017): bioinformatics. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-19-supplement-9.

Authors' contributions

TP and SK conceived and designed the research; SK and TP developed the method and were involved in drafting of manuscript; SK and SC implemented the software; JHY recruited liver cancer patients; YK generated MRM-MS data from the patients; SL designed the simulation studies. All the authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 08826, South Korea. [2]Department of Internal Medicine and Liver Research Institute, Seoul National University College of Medicine, Seoul 03080, South Korea. [3]Department of Biomedical Engineering, Seoul National University College of Medicine, Seoul 03080, South Korea. [4]Department of Mathematics and Statistics, Sejong University, Seoul 05006, South Korea. [5]Department of Statistics, Seoul National University, Seoul 08826, South Korea.

Published: 13 August 2018

References

1. Asnacios A, Fartoux L, Romano O, Tesmoingt C, Louafi S, Mansoubakht T, Artru P, Poynard T, Rosmorduc O, Hebbar M. Gemcitabine plus oxaliplatin (GEMOX) combined with cetuximab in patients with progressive advanced stage hepatocellular carcinoma. Cancer. 2008;112(12):2733–9.
2. DeVita VT Jr, Lawrence TS. DeVita, Hellman, and Rosenberg's Cancer: principles & practice of oncology, vol. 10e; 2009. p. 696–714.
3. Ferenci P, Fried M, Labrecque D, Bruix J, Sherman M, Omata M, Heathcote J, Piratsivuth T, Kew M, Otegbayo JA. Hepatocellular carcinoma (HCC): a global perspective. J Clin Gastroenterol. 2010;44(4):239–45.

4. carcinoma A-PWPoPoH. Prevention of hepatocellular carcinoma in the Asia–Pacific region: consensus statements. J Gastroenterol Hepatol. 2009;25:657–63.
5. Lin S, Hoffmann K, Schemmer P. Treatment of hepatocellular carcinoma: a systematic review. Liver Cancer. 2012;1(3–4):144–58.
6. Villanueva A, Toffanin S, Llovet JM. Linking molecular classification of hepatocellular carcinoma and personalized medicine: preliminary steps. Curr Opin Oncol. 2008;20(4):444.
7. Visser H, le Cessie S, Vos K, Breedveld FC, Hazes JM. How to diagnose rheumatoid arthritis early: a prediction model for persistent (erosive) arthritis. Arthritis Rheum. 2002;46(2):357–65.
8. Spitz MR, Etzel CJ, Dong Q, Amos CI, Wei Q, Wu X, Hong WK. An expanded risk prediction model for lung cancer. Cancer Prev Res. 2008;1(4):250–4.
9. Huang C-L, Liao H-C, Chen M-C. Prediction model building and feature selection with support vector machines in breast cancer diagnosis. Expert Syst Appl. 2008;34(1):578–87.
10. Gillette MA, Carr SA. Quantitative analysis of peptides and proteins in biomedicine by targeted mass spectrometry. Nat Methods. 2013;10(1):28–34.
11. Whiteaker JR, Lin C, Kennedy J, Hou L, Trute M, Sokal I, Yan P, Schoenherr RM, Zhao L, Voytovich UJ. A targeted proteomics-based pipeline for verification of biomarkers in plasma. Nat Biotechnol. 2011;29(7):625–34.
12. Kennedy JJ, Abbatiello SE, Kim K, Yan P, Whiteaker JR, Lin C, Kim JS, Zhang Y, Wang X, Ivey RG. Demonstrating the feasibility of large-scale development of standardized assays to quantify human proteins. Nat Methods. 2014;11(2):149–55.
13. Kline RB. Principles and practice of structural equation modeling. New York: Guilford publications; 2015. p. 7-24.
14. Hwang H, Takane Y. Generalized structured component analysis. Psychometrika. 2004;69(1):81–99.
15. Lee S, Choi S, Kim YJ, Kim B-J, Hwang H, Park T. Pathway-based approach using hierarchical components of collapsed rare variants. Bioinformatics. 2016;32(17):i586–94.
16. Llovet JM, Ricci S, Mazzaferro V, Hilgard P, Gane E, Blanc J-F, de Oliveira AC, Santoro A, Raoul J-L, Forner A. Sorafenib in advanced hepatocellular carcinoma. N Engl J Med. 2008;359(4):378–90.
17. Cheng A-L, Kang Y-K, Chen Z, Tsao C-J, Qin S, Kim JS, Luo R, Feng J, Ye S, Yang T-S. Efficacy and safety of sorafenib in patients in the Asia-Pacific region with advanced hepatocellular carcinoma: a phase III randomised, double-blind, placebo-controlled trial. Lancet Oncol. 2009;10(1):25–34.
18. Society AC: Cancer facts & figures 2017. American Cancer Society journal, CA: A Cancer Journal for Clinicians. 2017;17-18.
19. Kim H, Yu SJ, Yeo I, Cho YY, Lee DH, Cho Y, Cho EJ, Lee J-H, Kim YJ, Lee S. Prediction of Response to Sorafenib in Hepatocellular Carcinoma: A Marker Panel by Multiple Reaction Monitoring-Mass Spectrometry. Mol Cell Proteomics. 2017;mcp. M116:066704.
20. Lencioni R, Llovet JM. Modified RECIST (mRECIST) assessment for hepatocellular carcinoma. In: Seminars in liver disease: © Thieme Medical Publishers. Stuttgart, Germany; 2010. p. 052–60.
21. Chambers AG, Percy AJ, Simon R, Borchers CH. MRM for the verification of cancer biomarker proteins: recent applications to human plasma and serum. Expert Rev Proteomics. 2014;11(2):137–48.
22. MacLean B, Tomazela DM, Shulman N, Chambers M, Finney GL, Frewen B, Kern R, Tabb DL, Liebler DC, MacCoss MJ. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics. 2010;26(7):966–8.
23. Hwang H. Regularized generalized structured component analysis. Psychometrika. 2009;74(3):517–30.
24. Green PJ. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. J R Stat Soc Ser B Methodol. 1984:149–92.
25. McCullagh P. Generalized linear models. Eur J Oper Res. 1984;16(3):285–92.
26. Jang J-Y, Park T, Lee S, Kim Y, Lee SY, Kim S-W, Kim S-C, Song K-B, Yamamoto M, Hatori T. Proposed nomogram predicting the individual risk of malignancy in the patients with branch duct type Intraductal papillary mucinous neoplasms of the pancreas. Ann Surg. 2016;266(6):1062-8.
27. Team RC. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2014.
28. Gray J, Chattopadhyay D, Beale GS, Patman GL, Miele L, King BP, Stewart S, Hudson M, Day CP, Manas DM. A proteomic strategy to identify novel serum biomarkers for liver cirrhosis and hepatocellular cancer in individuals with fatty liver disease. BMC Cancer. 2009;9(1):271.
29. Braconi C, Meng F, Swenson E, Khrapenko L, Huang N, Patel T. Candidate therapeutic agents for hepatocellular cancer can be identified from phenotype-associated gene expression signatures. Cancer. 2009;115(16):3738–48.
30. Kong L-Q, Zhu X-D, Xu H-X, Zhang J-B, Lu L, Wang W-Q, Zhang Q-B, Wu W-Z, Wang L, Fan J. The clinical significance of the CD163+ and CD68+ macrophages in patients with hepatocellular carcinoma. PLoS One. 2013;8(3):e59771.
31. Tseng GC, Cheng C, Yu YP, Nelson J, Michalopoulos G, Luo J-H. Investigating multi-cancer biomarkers and their cross-predictability in the expression profiles of multiple cancer types. Biomark Insights. 2009;4:57.
32. Hutcheson J, Bourgo RJ, Balaji U, Ertel A, Witkiewicz AK, Knudsen ES. Retinoblastoma protein potentiates the innate immune response in hepatocytes: significance for hepatocellular carcinoma. Hepatology. 2014;60(4):1231–40.
33. Kinoshita M, Miyata M. Underexpression of mRNA in human hepatocellular carcinoma focusing on eight loci. Hepatology. 2002;36(2):433–8.
34. Cai Z, Daescu O, Li M. Bioinformatics research and applications: 13th international symposium, ISBRA 2017. Honolulu: Springer; 2017. May 29–June 2, 2017, proceedings, vol. 10330